

unpublished

PRELIMINARY COMBINATORIAL PROBABILITY MODEL
FOR THE VOYAGER QUARANTINE PROBLEM (PHASE I-II-)

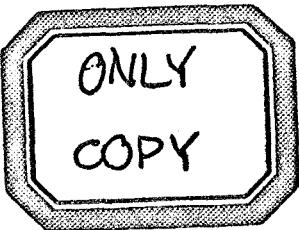


N 71-70687

FACILITY FORM 602	(ACCESSION NUMBER)	(THRU)
	10	
	(PAGES)	(CODE)
	CR-116909	(CATEGORY)
	(NASA CR OR TMX OR AD NUMBER)	

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

Sgt 64105 R



DOCUMENT NO. VOY-C2-TMI

October 1966

VOYAGER Master
Subject Index

PQ1A.F
0 50

Log Number

PRELIMINARY COMBINATORIAL PROBABILITY MODEL
FOR THE VOYAGER QUARANTINE PROBLEM (PHASE I-II-III)

16N10.051
y/m ser.no.

BY
T. F. Green

APPROVED

R. P. Wolfson
R. P. Wolfson, Cognizant Engineer
Planetary Quarantine
Voyager Spacecraft System Project

PREPARED FOR

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, California

Under JPL Contract No. 951112,
Modification No. 3

GENERAL ELECTRIC

Missile and Space Division
Valley Forge Space Technology Center
P. O. Box 8555, Philadelphia 1, Penna.

GENERAL ELECTRIC
MISSILE AND SPACE DIVISION
KIRKSVILLE

PROGRAM INFORMATION REQUEST/RELEASE

CLASS. LTR.	OPERATION	PROGRAM	SEQUENCE NO.	REV. LTR.
PDR NO.	—	5540-39	—	—

"USE "C" FOR CLASSIFIED AND "U" FOR UNCLASSIFIED

FROM

T. P. Green

TO

Distribution

DATE SENT

9/15/66

DATE INFO. REQUIRED

PROJECT AND REQ. NO.

REFERENCE DIR. NO.

SUBJECT

**PRELIMINARY COMBINATORIAL PROBABILITY MODEL FOR THE
VOYAGER QUARANTINE PROBLEM (PHASE I - II - III)**

INFORMATION REQUESTED/RELEASED

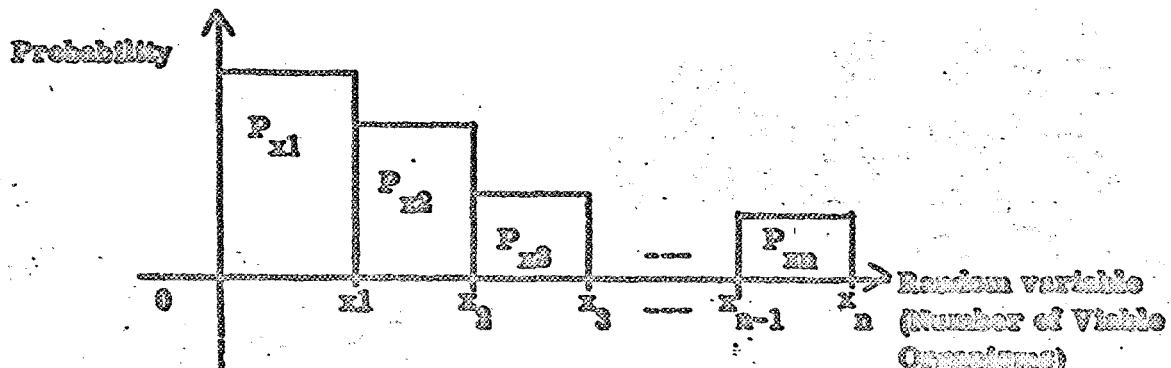
INTRODUCTION

The material in this writeup is a set of programming instructions for the development of a computing algorithm. This algorithm combines the probabilities of the survival of viable organisms of specified events from various sources. It is anticipated that this can be programmed on the G. E. D. S. C. S. for quick access capability. A similar version should be programmed on the large scale computer for later modification into the complete model to be developed for the determination of the probability of Martian contamination.

The model is based on a matrix whose columns represent events which could cause violation and whose rows represent independent sources of contamination. The construction of this matrix is described in the "Voyager Phase 1A, Task C, Bi-Monthly Report, Vol. 2". Some of the ideas leading to the model are the results of the combined efforts of several people, particularly R. P. Wilfzen and Dr. G. Ingram. In particular, the concept of interval probabilities and the associated calculation of the column probabilities were espoused by the latter. This was already tested on the D. S. C. S. and results have been satisfactory.

The working model computes the probability of each row and combines the results with those of the previous row recursively.

It has been concluded that the most efficacious way of presenting probabilities is by the "interval" concept. A graphical illustration appears below:



Distribution

F. S. Nayer (16) US141 VFSTC

V. Staub US605 VFSTC

R. Koenigsmill US605 VFSTC

T. P. Green -
Penn Park 0286

M. A. Martin -

4214 VFSTC

PAGE NO.

1 OF 3

RELEASING REQUIREMENTS

EXCEPT FOR MAINTAIN FOR

1 mo. 2 yrs.

6 mos. 6 mos.

3 mos. 12 mos.

1 yr.

$$\text{Note that } \sum_{j=1}^n P_{x_j} = 1.$$

The random variable (number of viable organisms) will consist of n intervals consisting of a graduated scale. This scale will be referred to hereafter as the "grid". The graduations have not currently been decided upon; however, it is expected they will look much like the following:

0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	
0	1	2	3	4	5	6	7	8	9	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	

It is anticipated that the exact nature of the grid will not significantly effect the design of the numerical process.

INPUT

Input consists primarily of row and column information. For a row, the input consists of an interval probability distribution. For a particular row, the columns (cells) will be provided the necessary input related to the probability that viable organisms will survive the event.

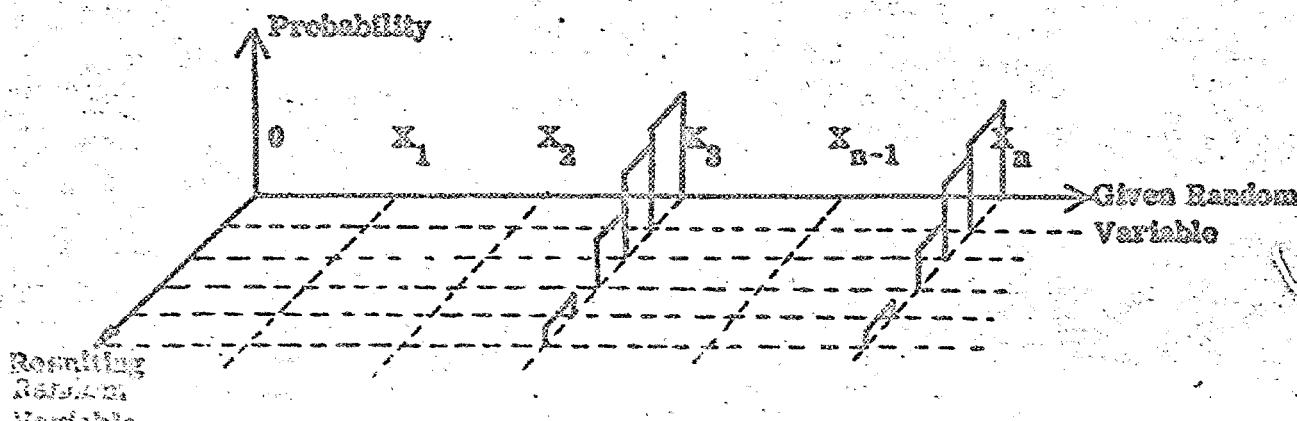
It has been decided that this input can be made in three ways which span the most likely ways that an experimenter can present his results.

(1) A simple proportion which describes the probability that one organism survives the event. This assumes that the event effects one organism independently of the other.

(2) An entire conditional distribution given in the interval probability form for various grid locations.

(3) An unconditional proportion which may be thought of as an approximation to (1) if only a weak probability statement can be made.

Each input will be described in detail. All three forms eventually result in a conditional distribution where the conditional variable is the given number of viable organisms. There will be a total of n of these distributions and they can be illustrated in the following diagram:



In the first, suppose the probability of an organism surviving is θ . Given $(x_{j-1} \rightarrow x_j)$ viable organisms the probability of $(x_{j-1} \rightarrow x_j)$ surviving is given by

$$\Pr\left\{x_{j-1} \leq x \leq x_j\right\} = \sum_{x=x_{j-1}+1}^{x_j} \binom{x_j}{x} \theta^x (1-\theta)^{x_j-x}$$

This generates all N conditional distributions.

Note that $\Pr\{0 \leq x \leq x_1\}$ given $(0 \leq x \leq x_1)$ is 1.

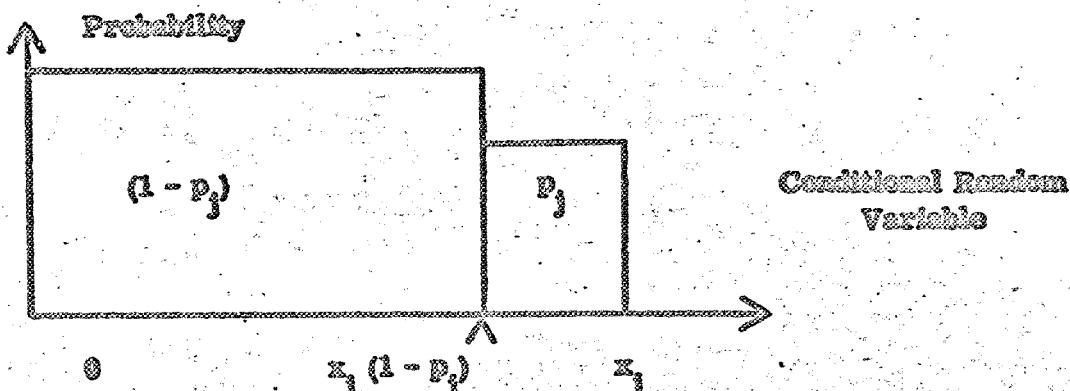
In the second, the entire conditional distribution is provided by the experimenter and assumes that he is completely knowledgeable about all values in the given grid.

In the third case, if it desired to associate a proportion surviving with various (or all) sample sizes in the given grid the following input option is provided:

<u>Given grid value</u>	<u>Proportion</u>
x_1	p_1
x_2	p_2
---	---
x_n	p_n

In many cases we may have $p_1 = p_2 = \dots = p_n$; that is the proportion is unconditional.

The conditional distribution is computed by assigning two probabilities $(1-p_j)$ and p_j to the intervals $(0 \rightarrow (1-p_j)x_j)$ and $((1-p_j)x_j \rightarrow x_j)$ resp. Thus the probability p_j is assigned to the expected value $x_j p_j$ viable organisms in the upper interval.



The grid is then introduced by proportioning both probabilities according to the interval sizes.

$$P_r \left\{ x_{j-1} \leq x \leq x_j \right\} \text{ given } (x_{j-1} \leq x \leq x_j) = \left(\frac{x_j - x_{j-1}}{x_j (1-p_j)} \right) (1-p_j)$$

If the interval is between $(0 \rightarrow x_j (1-p_j))$, and

$$P_r \left\{ x_{j-1} \leq x \leq x_j \right\} \text{ given } (x_{j-1} \leq x \leq x_j) = \left(\frac{x_j - x_{j-1}}{x_j p_j} \right) p_j$$

If the interval is between $(x_j (1-p_j) \rightarrow x_j)$.

In the event the interval encompasses the point $x_j (1-p_j)$ the probability is assigned proportionally:

$$P_r \left\{ x_{j-1} \leq x \leq x_j \right\} \text{ given } (x_{j-1} \leq x \leq x_j) = \left(\frac{x_j (1-p_j) - x_{j-1}}{x_j (1-p_j)} \right) (1-p_j) + \left(\frac{x_j - x_j (1-p_j)}{x_j p_j} \right) p_j$$

ROW PROBABILITY CALCULATIONS

The row calculations require the input source distribution and the conditional distribution at each cell provided by the three phase input options as described in the previous section. The "Voyager Phase 1A, Task C, Bi-Monthly Report No. 2" describes a way to progress from one cell to the other to construct the row probability. Since probabilities are now being described over intervals, this has to be somewhat modified although the basic formulas will not change. To illustrate, suppose the input row source probability is given by

<u>Interval</u>	<u>Probability</u>
$0 \rightarrow x_1$	p_1
$x_1 \rightarrow x_2$	p_2
•	•
•	•
$x_{n-1} \rightarrow x_n$	p_n

and the first cell contains the derived distribution.

	$0 \rightarrow x_1$	$x_1 \rightarrow x_2$	$x_2 \rightarrow x_3$	$x_3 \rightarrow x_4$		$x_{n-1} \rightarrow x_n$	Conditional Random Variable
$0 \rightarrow x_1$	p_{11}	p_{12}	p_{13}	p_{14}		p_{1n}	
$x_1 \rightarrow x_2$		p_{22}	p_{23}	p_{24}		p_{2n}	
$x_2 \rightarrow x_3$			p_{33}	p_{34}		p_{3n}	
$x_3 \rightarrow x_4$				p_{44}		p_{4n}	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$x_{n-1} \rightarrow x_n$						p_{nn}	

The probabilities p_1, \dots, p_n are then associated with the intervals across the top of the above table. The resultant probabilities (associated with the intervals along the left side) are computed by the following:

Interval	Probability
$0 \rightarrow x_1$	$\sum_{j=1}^n p_{1j} p_j$
$x_1 \rightarrow x_2$	$\sum_{j=2}^n p_{2j} p_j$
-----	-----
$x_{n-1} \rightarrow x_n$	$\sum_{j=n}^n p_{nj} p_j$

Appropriately, the probabilities along the left margin constitute the "marginal distribution of x . The marginal distribution probabilities are then associated with the top of the table for the next cell so that the calculations are then performed recursively from cell to cell. This whole process will be referred to as the "row probability calculation" or RPC.

COLUMN PROBABILITY CALCULATIONS

Once an EPC is completed, the result can be stored away and a new set of conditional probabilities calculated for the next row. The resultant probability is now associated with an additional source of viable organisms, so that the combinatorial probability will be calculated over the sum of the two.

Given the two distributions:

Interval	Probability	Interval	Probability
$0 \rightarrow x_1$	θ_1	$0 \rightarrow x_1$	ϵ_1
$x_1 \rightarrow x_2$	θ_2	$x_1 \rightarrow x_2$	ϵ_2
$x_2 \rightarrow x_3$	θ_3	$x_2 \rightarrow x_3$	ϵ_3
.....
$x_{k-1} \rightarrow x_k$	θ_n	$x_{k-1} \rightarrow x_k$	ϵ_n

The combinatorial probabilities are given by

Interval	Probability
$x_{k-1} \rightarrow x_1$	$\sum_{j=k}^n \frac{(x_H - x_L)}{(x_j + y_k - x_j - 1 - x_{k-1})} \theta_j \epsilon_k$

where

$$x_H = \text{smaller of } x_j \text{ and } x_{j-1} + x_k$$

$$x_L = \text{greater of } x_{j-1} \text{ and } x_j + x_{k-1}$$

$$\text{or } x_H - x_L = \theta_j x_j + \epsilon_k < x_{j-1} + x_{k-1} > x_1$$

The probabilities are multiplicative since the events are considered to be independent.

CALCULATION OF BINOMIAL PROBABILITIES.

Option (1) of the INPUT section provides for the calculation of binomial probabilities. The calculation involves excessively large quantities for increasing x_j .

For $(0 < x_j < 10^3)$ the recursion formula below is used:

$$Pr\{x+1\} = p_x \{x\} f(x).$$

where $f(x)$ is the ratio of the probability of the $(x+1)$ 'st term over the probability of the x 'st term and equals

$$\frac{\binom{x_j}{x+1} p^{(x+1)} (1-p)^{x_j-x-1}}{\binom{x_j}{x} p^x (1-p)^{x_j-x}} = \left(\frac{x_j-x}{x+1}\right) \left(\frac{p}{1-p}\right).$$

Thus the calculation of combinations is avoided.

To create interval probabilities, the

$$Pr\{0 < x \leq 1\} \text{ is set equal to } Pr\{x=1\}$$

and in general

$$Pr\{x_1 < x \leq x_1 + 1\} \stackrel{\Delta}{=} Pr\{x = x_1 + 1\}$$

where $\stackrel{\Delta}{=}$ is read as "defined as". This approximation is only used for intervals of length "1". Otherwise $Pr\{x_1 < x \leq x_{1+1}\} = Pr\{x \leq x_{1+1}\} - Pr\{x \leq x_1\}$.

For $(x_j > 10^3)$ the binomial can be approximated by the normal distribution provided

$$\left(\frac{1}{x_j + 1} < p < \frac{3}{x_j + 1} \right).$$

The approximation is

$$\Pr \left\{ x_i \leq x \leq x_{i+1} \right\} = \Phi \left(\frac{x_{i+1} + \frac{1}{2} - x_j p}{\sqrt{x_j p (1-p)}} \right) - \Phi \left(\frac{x_i - \frac{1}{2} - x_j p}{\sqrt{x_j p (1-p)}} \right)$$

Where Φ represents cumulative percentage points of the standard normal distribution.

These can be approximated by a process described in a T. I. S. titled "A Computer Algorithm for the Error Integral" by G. M. Roe T.I.S. # 66-C-050. Studies have shown this algorithm to be sensitive to probability changes as far away from the mean as 11 or 12 standard deviations. These were made by R. Wharton of the Engineering and Scientific Computer Applications Group in FACMO, MSD. The algorithm has already been placed on the auxiliary library for our computer systems (D. S. C. S. and main).

After interval probabilities have been calculated, the results should be normalized by

$$P_X \left\{ x_i \leq x \leq x_{i+1} \right\} = \frac{\Pr \left\{ x_i \leq x \leq x_{i+1} \right\}}{\sum \Pr \left\{ x_i \leq x \leq x_{i+1} \right\}}$$

to preserve the basic necessary condition that all probabilities should sum to unity.

Approximations in the low interval regions combined with numerical loss of significance plus other approximations will no doubt disturb the condition somewhat.

FUNCTIONAL FLOW DIAGRAM

N , number of intervals in basic grid.

Given: NR, number of rows.

NC, number of columns.

Set k = 1

1. Read in basic grid pattern.

2. Read in source probability distribution for row k.

Grid	Probability
$0 \rightarrow x_1$	p_1
$x_1 \rightarrow x_2$	p_2
-----	-----
$x_{n-1} \rightarrow x_n$	p_n

3. Set $I = 1$
4. Read in probability option for column I.
5. Construct conditional probability distribution according to either
 - option (1): given p , use binomial routine.
 - option (2): entire distribution.
 - option (3): unconditional proportions.
6. Perform row probability combinatorial algorithm to construct marginal input probabilities for column $I + 1$.
7. Set $I = I + 1$
8. If $I > NC$ call column combinatorial algorithm subroutine to combine the probabilities of row k to the previous row resultant.
9. Set $k = k + 1$
10. If $k \leq NR$, recycle back to step 2 to perform another row calculation. If $k > NR$ print the resultant probability density:

Grid	Probability
$0 \rightarrow x_1$	p_1
$x_1 \rightarrow x_2$	p_2
-----	-----
$x_{n-1} \rightarrow x_n$	p_n